CDS Project Proposal: Roivant Sciences NER project

Kailing Wang (kailing.wang@nyu.edu)

What is the project?

The project is to establish the document analysis pipeline where given a list of targets of interest, we use/improve our Named Entity Recognition (NER) to find drug names that we don't already have in our drug name dictionary from millions of medical papers. The company has made previous efforts on this problem, that is building a deep neural net NER purely based on the syntax of documents. In this project, we will leverage ChEMBL dataset, which serves as our training set, taking advantage of the sentence structures from those reference documents. We would expect that this additional assistance/information from ChEMBL could expand our named entity recognition (NER) capabilities. This methodology could be easily generalized/applied into other scenarios –provided a list of entities, making recommendations of their relevant entities from unstructured data.

What does the data look like?

The dataset is ChEMBL(https://en.wikipedia.org/wiki/ChEMBL), which is a manually curated medical database. An NDA will be required. The data is cleaned and accessible from our MySQL database.

ChEMBL dataset:

	A	B	C	D	E	F	G
1	target_pref_name	target_synonyms	drug_pref_name	drug_synonyms	ref_type (mechanism_refs)	ref_url (mechanism_refs)	action_type
791099	Human herpesvirus 1 DN	2.7.7.7	PENCICLOVIR	Fenistil	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=8d119057-f299-43ea-b516-e84a09cab890	INHIBITOR
791100	Human herpesvirus 1 DN	3.1.26.4	PENCICLOVIR	Fenistil	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=8d119057-f299-43ea-b516-e84a09cab890	INHIBITOR
791101	Mu opioid receptor	MOR1	OXYCODONE	Oxicone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791102	Mu opioid receptor	OPRM1	OXYCODONE	Oxicone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791103	Mu opioid receptor	Mu-type opioid receptor	OXYCODONE	Oxicone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791104	Mu opioid receptor	Mu opiate receptor	OXYCODONE	Oxicone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791105	Mu opioid receptor	Mu opioid receptor	OXYCODONE	Oxicone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791106	Mu opioid receptor	M-OR-1	OXYCODONE	Oxicone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791107	Mu opioid receptor	MOR-1	OXYCODONE	Oxicone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791108	Mu opioid receptor	MOP	OXYCODONE	Oxicone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791109	Mu opioid receptor	hMOP	OXYCODONE	Oxicone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791110	Mu opioid receptor	MOR1	OXYCODONE	Oxycodone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791111	Mu opioid receptor	OPRM1	OXYCODONE	Oxycodone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791112	Mu opioid receptor	Mu-type opioid receptor	OXYCODONE	Oxycodone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791113	Mu opioid receptor	Mu opiate receptor	OXYCODONE	Oxycodone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791114	Mu opioid receptor	Mu opioid receptor	OXYCODONE	Oxycodone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791115	Mu opioid receptor	M-OR-1	OXYCODONE	Oxycodone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791116	Mu opioid receptor	MOR-1	OXYCODONE	Oxycodone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791117	Mu opioid receptor	MOP	OXYCODONE	Oxycodone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791118	Mu opioid receptor	hMOP	OXYCODONE	Oxycodone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791119	Mu opioid receptor	MOR1	OXYCODONE	Oxycontin	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791120	Mu opioid receptor	OPRM1	OXYCODONE	Oxycontin	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791121	Mu opioid receptor	Mu-type opioid receptor	OXYCODONE	Oxycontin	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791122	Mu opioid receptor	Mu opiate receptor	OXYCODONE	Oxycontin	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791123	Mu opioid receptor	Mu opioid receptor	OXYCODONE	Oxycontin	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791124	Mu opioid receptor	M-OR-1	OXYCODONE	Oxycontin	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791125	Mu opioid receptor	MOR-1	OXYCODONE	Oxycontin	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791126	Mu opioid receptor	MOP	OXYCODONE	Oxycontin	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791127	Mu opioid receptor	hMOP	OXYCODONE	Oxycontin	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791128	Mu opioid receptor	MOR1	OXYCODONE	Proladone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791129	Mu opioid receptor	OPRM1	OXYCODONE	Proladone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791130	Mu opioid receptor	Mu-type opioid receptor	OXYCODONE	Proladone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791131	Mu opioid receptor	Mu opiate receptor	OXYCODONE	Proladone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791132	Mu opioid receptor	Mu opioid receptor	OXYCODONE	Proladone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791133	Mu opioid receptor	M-OR-1	OXYCODONE	Proladone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791134	Mu opioid receptor	MOR-1	OXYCODONE	Proladone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791135	Mu opioid receptor	MOP	OXYCODONE	Proladone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791136	Mu opioid receptor	hMOP	OXYCODONE	Proladone	DailyMed	http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=dd9f9d6d-9a59-4f74-8fe7-685e0c3c5fd4#nlm34(	AGONIST
791137							
791138							
791139							

chembl_drug_target_publicationv +

The raw dataset contains nearly 800,000 records. However, note that a large proportion of the dataset is the combinations of various drug aliases and target aliases. So if we remove those duplicates, there remains 7,320 records. 20% of the data will be reserved for testing.

- Step1: Preprocessing

From ChemBl list of sentences, extract two types of predicates based on the outcome of Syntaxnet's dependency parser and POS tagging: (1) “Good Predicates”: predicates that link a target to a drug; (2) “Other Predicates”: predicates that link a target to a word that is not a drug name. For example,

1	target_pref_name	target_synonyms	drug_pref_name	drug_synonyms	ref_type (mechanism_refs)	ref_url (mechanism_refs)	action_type
155404	Cyclooxygenase	COX1	PIROXICAM	CP-16171	PubMed	http://europepmc.org/abstract/MED/11153163	INHIBITOR

Abstract

Non-opioid analgesics are some of the most widely used therapeutic agents in clinical practice today. The number of patients at risk for adverse events related to the use of these agents is rapidly expanding. While the gastrointestinal toxicity of these medications is well known, it has become increasingly apparent that the kidney is also an important target for untoward clinical events. Evidence of the nephrotoxicity of analgesic preparations is not sufficiently completed and available in our region. Analgesic-related renal injury has been classified based on mechanism of action into "classic" analgesic nephropathy and NSAID-related renal toxicity. From clinical point of view the renal side effects induced by analgesics can be classified into hemodynamic (functional) side effects and idiosyncratic side effects. The common link in both types of side effects seems to be renal ischemia related to prostaglandin synthesis inhibition. Key enzyme in this process is cyclooxygenase occurring in two isoforms: COX-1 and COX-2. Antiinflammatory effect of NSAIDs is mediated by COX-2 inhibition, while the side effects (gastrotoxicity, nephrotoxicity) by inhibition of COX-1. COX-1 was more inhibited by indomethacin and piroxicam and COX-2 by 6-MNA (active metabolite of nabumetone), diclofenac and ibuprofen. Nimesulide and meloxicam selectively block COX-2 and are recommended to patients at risk or treated with diuretics. (Tab. 2, Fig. 2, Ref. 38.)

For the 155404th record, we could extract “was more inhibited by” as a “Good Predicate”.

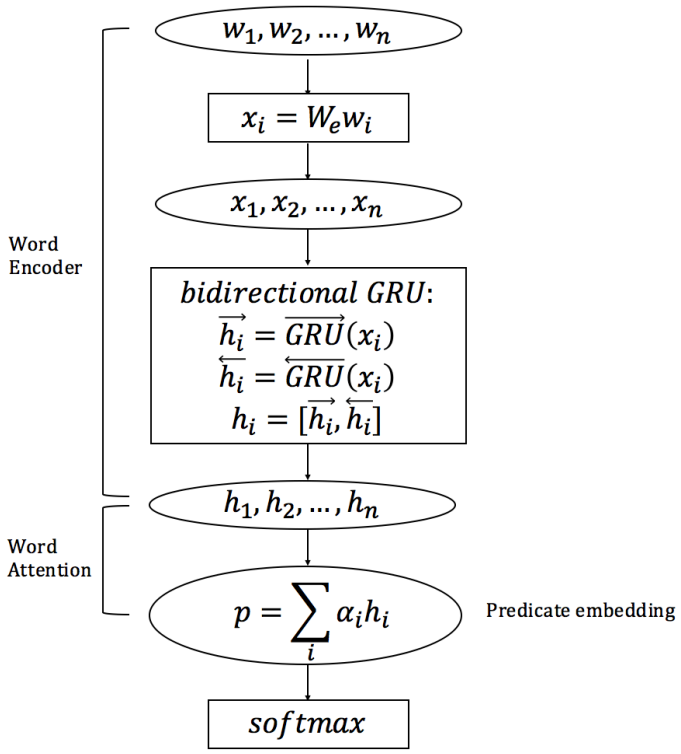
- Step 2: Training a predicate classifier based on Hierarchical Attention Network (HAN)[1]

Take 80% “Good Predicates” and 80% “Other Predicates” as the training set. The HAN-based classifier consists of several parts: a word sequence encoder (using a pre-trained word embedding and a bidirectional GRU), a word attention layer and the softmax. The other 20% “Good Predicates” and 20% “Other Predicates” are used for validation.
The classifier differentiating “Good Predicates” and “Other Predicates” is for reducing false positive errors in the final model, and finding the optimum embedding representation of a predicate (using HAN to identify the most important words from a predicate).

- Step 3: Testing

Firstly, locate the target name at a sentence level on test set; Secondly, if the predicate in the testing sentence is determined as a “Good Predicate” by the classifier, the other noun phrase in the testing sentence will be extracted as a new drug name which acts on the given target.

[1] https://www.cs.cmu.edu/~hovy/papers/16HLT-hierarchical-attention-networks.pdf



- Step 4: Improvement - Evaluate the need for drug NER to improve efficiency /performance

  Firstly, use our current NER to identify drug names; Secondly, apply the classifier to validate whether the drugs recognized are truly acting on the given targets.

What are the rubrics of success?

Evaluation metrics:

(1) $\text{Recall} = \frac{\text{the number of drug names an NER correctly detected}}{\text{the total number of drug names contained in the input text}}$

(2) $\text{Precision} = \frac{\text{the number of drug names an NER correctly detected}}{\text{the total number of drug names identified by the NER}}$

(3) $\text{F} - \text{measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

The following NER methods could serve as baselines/comparisons:

1. simple Tagging model
2. our current NER (based on the syntax of documents)
3. replacing word-embedding with fuzzy matching in Step 2: https://pypi.python.org/pypi/fuzzywuzzy

What is the relevant organizational, project or institutional history ?

Roivant Sciences is a biopharmaceutical company focused on completing the development of promising late-stage drug candidates. There are many challenges involved in identifying promising inlicensing opportunities, including:

1. there are tens of thousands of drugs across thousands of companies, with no centralized complete database.
2. due diligence for any opportunity happens across many categories of data, including drug efficacy, market size, IP, development cost, sales force size.
3. each of these categories of data has structured and unstructured data, at best incomplete and spread across multiple sources.

Roivant is in the process of building our methodology for extracting information of interest from unstructured data.

How will the organization support and mentor the students?

The instructors would be:

1. Bill McMahon <bill.mcmahon@roivant.com>, VP of Computational Research at Roivant Sciences
2. Yoann Mamy Randriamihaja <yoann.randriamihaja@roivant.com>, director of Computational Research at Roivant Sciences